自然言語処理 125－9
（1998. 5. 29）

基本レベルカテゴリに基づいた名詞語彙知識の獲得

河部 恒, 松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{kou-k,matsu}@is.aist-nara.ac.jp

自然言語処理において辞書の存在は不可欠である。現在いろいろな種類の機械可読辞書が、システムと独立したモジュールとして幅広い分野で使われている。
自然言語処理の意味処理を行おうとするとき、動詞に関しては、それが述語の働きをするという観点から、格フレームが用いられており、それを基に動詞辞書がつくられている。しかし名詞については、どのような辞書記述が必要かについての共通の構造がない。
それは、名詞に関しての辞書をつくろうとした場合、多義性、世界知識の問題等、様々な困難が存在するためである。本論文では、その原因の一つとして、すべての名詞を機能的に分類することをしないですべて統一的の扱おうとしていたことに起因するという視点を導入し、名詞を類別することを考える。特にシソーラスのように名詞間の階層関係を決定する場合には、上位下位関係は各レベルで均等ではなく、ある特徴を持ったレベル (=基本レベルカテゴリ) が存在すると仮定し、その選定を行った。さらにそこで得られたカテゴリに基づいて名詞をクラス分けすることを試みた。

[キーワード] 辞書, コーパス, 語彙知識, シソーラス

Acquisition of Nominal Lexical Knowledge based on Basic Level Category

KAWABE Kou and MATSUMOTO Yuji
Graduate School of Information Science, Nara Institute of Science and Technology

Dictionaries are one of the important resources for Natural Language Processing(NLP).
In semantic processing of NLP, verbs are normally represented in the form of case frames because they are considered to function as predicates. A verb dictionary has been constructed based on that idea. Nouns, however, cannot be analyzed in such a way.
In this thesis, we address that one of the reasons are caused by the idea that regards all the representation of nouns as the same. Therefore, we try to classify nouns into several categories. Especially, in the case of deciding the hierarchy of nouns, like a thesaurus, we assume that each level of the hierarchy is not equivalent and a certain level is characteristic, which is called the basic level category. We investigate a way to identify them.

[keyword] Lexicon, Corpus, Lexical Knowledge, Thesaurus

1 はじめに

辞書は自然言語処理には欠かすことのできない言語資源のひとつであり、解析の多くの段階で単独のモジュールとして扱われる。辞書を構築するのは多大の人手と時間を必要とするので、もし辞書を汎用的に記述することができればコストを大幅に削減することができ、また記述の一貫性という観点からも望ましい。

現在の辞書は、人間用の辞書、機械可読辞書、シソーラス、概念辞書、格フレーム辞書などがあり [9, 10, 2]、またLFGの語彙項目や、HPSGのSEM素性も広い意味で辞書ととらえることができる。

たとえば動詞については、それが述語の働きをするという観点から、格フレームや概念構造 [1] といった記述方法が考えられている。これらは深層的構造をもとに動詞の意味を分類するということに成

上位レベル	植物	家具
基本レベル	花	椅子
下位レベル	バラ	肘掛け椅子

表 1: 基本レベルと上位下位階層

功している。

しかし、名詞については語彙とその意味をどのように記述したらいいかについてのコンセンサスが今のところ得られていない。たとえば長い漢字複合語や名詞接続助詞「の」による“AのB”といった表現は形態素/構文解析ではこれ以上解析できず、このままでは形態素/単語のあいだの関係を推し量ることができない。

本研究では、従来のように名詞を機能的に分類せず、すべて統一的に扱おうとしていることにその原因があると考える。名詞の間でも人間にとってより身近なものとそうでないものがあるという考え方[3]に基づきまずそれらを選別することを試みた。ここでは認知言語学の基本レベルカテゴリという概念を援用し、最終的にそれを元に辞書を構築することを目的とする。語彙はまずこのレベルで記述され、これより下位の語彙は基本レベルカテゴリの性質の多くを継承することができ、記述量の削減が期待できる。

以下の節では、まず 2 節で本研究で重要な概念となる基本レベルカテゴリ (BLC) について説明する。次に 3節で BLC をその性質に基づいてコーパスから抽出する方法を述べる。4節で実験の結果を示し、5節でまとめを述べる。

2 基本レベルカテゴリとは

この節では、基本レベルカテゴリとはなにかについて説明し、その性質について述べる。

カテゴリとは、人間が世界を認識する上で何らかの属性や性質を共有する集まりのことである [3]。従来の考え方では、属性を持つか否かによってその境界ははっきり区別されるとして扱われてきた。しかし、たとえば“スポーツ”という単語を考えた場合、それが示すものは様々な性質を持ち、それらすべてに共通の属性を挙げることは難しい。それらはある部分重なり合い全体としてスポーツと呼ばれるのである (図 2)。このような、すべてに共通な属性はないがお互いは似ている関係を家族的類似性 [8] と呼ぶ。この時カテゴリの成員の中でよりスポーツらしいか、そうでないかといった違いが生じる。

基本レベルカテゴリとは、このようなカテゴリの上位下位関係を考えた場合、人間が理解する上でもっとも“丁度良い”レベルのことである。表 2 は、植物→花→バラ および 家具→椅子→肘掛け椅子という IS-A 階層を表しているが、“植物”のレベルでは、対象が抽象的過ぎて植物一般に共通する属性を挙げることは難しいし、“バラ”のレベルでは逆に特殊すぎる。“花”のレベル、すなわち基本レベルが人間が理解する上で丁度良いレベルとなっている。

Rosch ら [7] によると

- カテゴリーの呼び名としてもっとも一般的に使用されるレベル
- もっとも短い基本的な語彙素レベル
- 子供によって理解される最初のレベル
- 我々の知識のおおかたが組織化されるレベル
- 語が中立的なコンテキストで用いられるレベル

等が基本レベルカテゴリの性質としてあげられている。

知識のおおかたが組織化されるレベルが基本レベルであるならば、そのレベルに注目して語彙知識を書いておけば、それより下位のレベルの語彙は基本レベルの性質の多くを継承することが考えられる。(図 1) また基本レベルより下位の成員間の類似度は高く、基本レベル同士の類似度は低いことが予想されるので、基本レベルという粒度で記述するということは冗長さのより少ない記述が可能であると予想される。以上の理由から、基本レベルに注目しその語彙知識の記述を目指すことが妥当であると考えられる。そこで、次ではこの基本レベルカテゴリの名詞をコーパスから自動的に抽出する方法について述べる。

3 抽出方法

3.1 抽出に利用した BLC の性質

基本レベルカテゴリの名詞を抽出する際に、上記2節にあげた性質のうちコーパスにあらわれる以下の 3 つの手がかりを用いた。

- A 頻度が高い

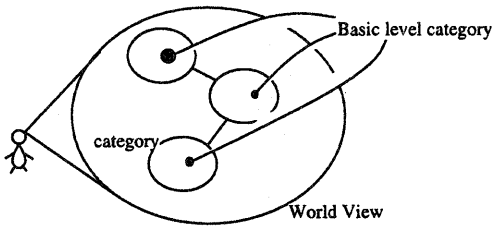

図 1: World View and Basic Level Category

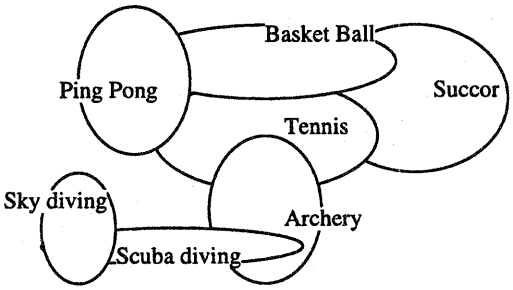

図 2: 家族的類似性 (スポーツ)

名詞	頻度	その	この	あの
植物	1004	0	0	0
花	9914	23	39	0
バラ	2910	0	0	0

表 2: 1994 年度毎日新聞一年分における頻度

- B 照応表現に用いられやすい。

  表 2 は、新聞一年分における “植物”,“花”,“バラ” の総頻度と「その + “植物”」 のような照応表現の頻度をあらわしたものである。“植物” と “バラ” は一度も照応されていないが、基本レベルである “花” では計 62 回あらわれている。これは基本レベルカテゴリの名詞が照応されるときの言い換え表現に用いられやすいことをあらわしている。

- C 名詞複合語の構成要素になりやすい。

  - 植物束 ?
  - 花束 ○
  - バラ束 ?

コーパス	異なり語彙数	総出現回数
毎日新聞 91 年度	34,968	5,781,144
92 年度	35,362	5,983,755
93 年度	35,460	5,418,550
94 年度	37,515	6,366,202
日経新聞 (経済分野)	10,701	282,974
(政治分野)	11,228	328,475
森鴎外小説	8,262	76,108
夏目漱石小説	7,257	58,195

表 3: コーパスの性質

のような例を考えると、上位や下位の名詞では構成要素になりにくいという性質がみられる。逆に基本レベルの名詞は複合語の構成要素、特に最後の要素になりやすい。

3.2 コーパス

コーパスは、新聞と小説を対象とした。まず第一に毎日新聞の 1991 年から 1994 年 (各年 90 - 100 Mbyte 程度)、次に分野の違いにより基本レベルカテゴリがどう変化するかをみるために日経新聞 1994 年の経済分野と政治分野の 2 種類 (計 10Mbyte 程度)、最後に小説として、電子化されている夏目漱石、森鴎外 (計 2Mbyte 程度) を使用した。表 3 にコーパスの性質を示す。

3.3 アルゴリズム

以下に抽出のためのアルゴリズムを示す。抽出方法は利用した 3 つの性質それぞれに対して 3 通りあり、それぞれ A, B, C とする。

3.3.1 総頻度 (A)、照応詞頻度 (B) による方法

方法 A,B はほぼ同じなのでまとめて説明する。

1. コーパスに品詞タグをふる。[1]
2. 普通名詞を取りだしカウントする。(=総頻度 c) ただし 頻度 C 以下のものは取らない。ここで言う普通名詞とは 表 4 の分類で言うところの普通名詞にあたる。

[1] 新聞に対してはすでに RWCP のタグがついていたのでそれを利用し、小説に対しては茶筅 [11] を用いた。

3. “その”, “この”, “あの” に続く普通名詞を取りだしカウントする。(=照応詞頻度 r) ただし、頻度 R 以下, 照応率 $RC(=\frac{r}{c})$ 以下のものはとらない。

4. 取り出された名詞に対して分類語彙表 [9] を引く。

5. 上位5桁までみてグループ分けをし、その中から以下の基準で取り出す。

   - 人工物 (1.4), 自然物 (1.5) に属するもの
   - 総頻度が最大のもの (実験 A)
   - 照応詞頻度が最大のもの (実験 B)

以下で、各 step について解説する。

step 2. で普通名詞のみを対象にしているのは、固有名詞や数詞といった種類の名詞は基本レベルカテゴリとは考えにくいからである。またここでは動詞との関連を考えにいれないとしサ変名詞を除いている。

step 3. で照応する単語として連体詞 “その”, “この”, “あの” の3種類を選んだ理由は、この3種類がコーパス中でもっとも頻度が高かったからである。

step 5. で分類語彙表を採用しているのは、得られた多数の基本レベルカテゴリ候補の名詞を一定の基準で選別するためである。コーパスから抽出するとき、カテゴリ候補の名詞の集合は、ある特定のカテゴリに多く集まり、また別のカテゴリにはほとんどないということが起り得る。よって単に頻度だけで選ぶのでは得られた基本レベルカテゴリのセットにばらつきが生じてしまう。そこで分類語彙表で一定レベルより細かいところは同一視することにより足切りしていったんグループ化し、そこからもっとも頻度/照応詞頻度の高いものを選ぶことで得られる候補のカテゴリ間の片寄りをなるべく少なくするという働きをしている。なお、1.4 と 1.5 に属するもののうち上位5桁を同一視したときの分類語彙表は 136 個のグループに分かれる。

3.3.2 複合語の構成要素 (C) による方法

方法 C では、複合語の構成要素になりやすいという性質から基本レベルカテゴリの名詞を取り出す。ここでは対象データとしてコーパスではなく、辞書の見出し語 (18万語) を利用した。それには以下の理由がある。

普通名詞
サ変名詞
固有名詞
地名
人名
数詞
形式名詞
副詞的名詞
時相名詞

表 4: 名詞の分類

- コーパスから取り出した場合に生じるような切り出し誤りがない。
- 頻度の非常に少ない単語も取り出せる。

また、あまり長い形態素は基本レベルカテゴリになりにくく、また漢字1字の場合は、“性” や “的” といった接辞が圧倒的なのでこれらを省いた。結局、見出し語のすべての部分文字列[2]のうち、

- 長さが 2 から 4 文字で、
- 漢字のみからなり、
- 頻度の高いもの

を基本レベルカテゴリの候補とした。

4 実験結果

この節では実験結果について述べる。毎日新聞から取り出された名詞の数は 1994 年度の場合で表4、5 のようになった。ただし、c = 総頻度、r = 照応詞頻度、$RC(=\frac{r}{c})$ =照応率 である。

照応率 $\frac{r}{c}$ が上がるにしたがって対象となる名詞は急速に減っていくことが分かる。実験で使ったパラメータは新聞記事では $C \geq 100$, $R \geq 1$, $RC > 0$, 小説 では $C \geq 1$, $R \geq 1$,$RC > 0$, とした。

4.1 方法 A,B

方法 A,B による結果では以下のようになった。

- 新聞記事では 4 年分で各年同じようなものが取り出せた。
  89/136 個が共通

[2]長さ n の漢字列で $\frac{n*(n+1)}{2}$ 種類の部分文字列がある。

表 5: r≥1

c	r/c=0	r/c=0.01	r/c=0.05
10	20324	3043	625
50	10541	2036	164
100	7119	1220	107
300	3437	586	62

表 6: r≥2

c	r/c=0	r/c=0.01	r/c=0.05
10	3128	1695	350
50	2911	1478	164
100	2648	1215	107
300	1904	586	62

- 小説では芥川竜之介と森鴎外ではかなり異なるものが取り出せた。
  25/136 個が共通

- 小説と新聞で共通に取り出せたものは以下の9つ。
  傘、部屋、本、車、金、風、土地、手、死

以下は結果の一部例である。数字は分類語彙表の分類番号上位5桁をあらわし、次が分類語彙表についている見出しである。そのレベルに存在する名詞のならびのうち、下線のついているのが基本レベルカテゴリとして今回抽出された名詞である。

- 1.4660 乗り物 (海上)
  船、漁船、ボート、空母、タンカー、船舶、艇

- 1.4640 計器
  時計、腕時計、カウンター、メジャー、計

- 風
  風、台風、強風、追い風、あらし、緑風

政治面と経済面によるドメインの違いによる実験は以下のような差が見られた。(方法 B での上位20個による結果)

- 政治面、経済面で共通
  ところ、国、時期、場、中、点、日、背景、分、問題、理由

- 政治面のみ
  内閣、政権、政治、事件、方針、典型、過程、時点

- 経済面のみ
  反動、分野、種、経済、内容、成果、傾向、地域

4.2 方法 C

方法 C による結果の例を以下に示す。(上位 10 位)

大学, 主義, 日本, 天皇, 文字, 社会, 奉行, 運動, 神社, 太郎

4.3 考察

方法 A, B, C の結果の違いを述べる。まず総頻度をみる方法 A ではソースコーパスのドメインの影響を強く受けていた。たとえば、新聞記事からとった基本レベルカテゴリの名詞の候補は、"死亡", "官邸", "銀行" といったものであったのに対し、小説では、"僕", "自分", "女" といったものが目立った。次に、辞書の見出し語から複合語の構成要素を取り出す方法 C では、大きく分けて 2 種類の結果が得られた。第 1 に "部屋", "電気", "鉄道" といった、基本レベルとしてふさわしいもの、第 2 に "山椒魚", "白子" といったとても頻度の低いものである。またここでは漢字 1 字のものはあらかじめ除いてしまったので、"花" 等の本来基本レベルにはいるべきものも除かれてしまった。最後に照応頻度をみる方法 B であるが、この方法では "車", "部屋", "飛行機" といった基本レベルとするにふさわしいものが多く抽出された。

政治面と経済面のドメインの違いによる基本レベルの差は、約半分が共通、残り半分がドメイン独自の名詞という結果になった。基本レベルのカテゴリの名詞は、記述されている内容が細かくなればより下位のレベルにシフトするということも考えられる。たとえば、園芸家のドメインではもはや "花" は基本レベルではなく、もっと個別の名詞、たとえば "バラ" が基本レベルとなり、もっと下位の "蔓バラ" や "ミニバラ" について話し合われるということもありうる。

5 おわりに

本研究では、認知言語学の基本レベルカテゴリという考え方を元にして、名詞の語彙記述をすることを目的に、3つの方法を用いてその候補をコーパスから自動的に取り出す方法について述べた。辞書を構築する際に、ここで得られた基本レベルカテゴリの名詞をまず記述することで、記述量を押さえ、認知的にも妥当な語彙記述ができると考えられる。

現在は、コーパスからの手がかりに基づいて名詞がここで得られたカテゴリのどれの属すかを判定する(クラス分け)実験を行っている。また、実際に語彙を記述する方式については、たとえばGenerative Lexicon[6, 4, 5] のような feature に基づいて、動的に意味を導出していく枠組みを考えている。

参考文献

[1] R. Jackendoff. *Semantic structures.* MIT press, 1990.

[2] Betty Kirkpatrick. *Roget's Thesaurus of English words phrases.* Longman, 1987.

[3] George Lakoff. 認知意味論. 紀伊國屋書店, 1993.

[4] J. Pustejovsky, S. Bergler, and P. Anick. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2):331–358, 1993.

[5] J. Pustejovsky and B. Boguraev. Lexical knowledge representation and natural language procesing. *Artificial Intelligence*, 63(1-2):193–223, 1993.

[6] James Pustejovsky. *The Generative Lexicon.* The MIT Press, 1995.

[7] Rosch, Eleanor, C. Mervis, W. Gray, D. Johnson, and P. Boyes Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[8] Ludwig Wittgenstein. *Philosophical Investigations(*哲学的探求*).* Macmillan, 1953.

[9] 国立国語研究所. 分類語彙表（増補版）. 国立国語研究所, 1996.

[10] 日本電子化辞書研究所. *EDR* 電子化辞書使用説明書 *(*第 *2* 版*).* 日本電子化辞書研究所, 1995.

[11] 松本 裕治, 北内 啓, 山下 達雄, 今一 修, and 今村 友明. 日本語形態素解析システム『茶筌』version 1.0 使用説明書. Information Science Technical Report NAIST-IS-TR97007, Nara Institute of Science and Technology, 1997.